Claims

What is claimed is:

1. A method of processing a request to at least one server, comprising the steps of:

receiving a request; and

scheduling submission of the request to the at least one server based on: (i) a quality-of-service (QoS) class assigned to a client from which the request originated; (ii) a response target associated with the QoS class; and (iii) an estimated response time associated with the at least one server.

2. The method of claim 1, further comprising the step of withholding the request from submission to the at least one server when the request originated from a client assigned to a first QoS class to allow a request that originated from a client assigned to a second QoS class to meet a response target associated therewith.

3. The method of claim 2, further comprising the steps of:

determining a throughput of the at least one server; and

reducing a request withhold rate to increase throughput of the at least one server.

4. The method of claim 2, further comprising the steps of:

monitoring a throughput of the at least one server; and

varying a request withhold rate to balance the throughput and request response times.

5. The method of claim 1, further comprising the step of assigning the response target to the QoS class.

6. The method of claim 5, wherein the step of assigning the response target to the QoS class further comprises the step of assigning a response time target to the QoS class.

7. The method of claim 5, wherein the step of assigning the response target to the QoS class further comprises the step of assigning a response percentile target to the QoS class.

8. The method of claim 1, further comprising the step of estimating the response time associated with the at least one server based on one or more requests sent to the at least one server within a given time period.

9. The method of claim 1, further comprising the step of assigning a target response time to a plurality of QoS classes in which lower quality classes are assigned larger response times than higher quality classes.

10. The method of claim 1, further comprising the steps of:

determining dispatch times for requests from a difference between at least one predicted response time of the at least one server and the target response time corresponding to the QoS class of the request; and

sending requests to the at least one server based on dispatch times.

11. The method of claim 1, wherein a plurality of applications are running on the at least one server and requests are routed to applications, further comprising the steps of:

estimating response times of applications based on one or more requests sent to the applications within a time period; and

sending a request to an application whose estimated response time is not greater than a target response time corresponding to the QoS class of the request.

12. The method of claim 11, further comprising the step of varying a number of requests sent to applications so that estimated response times of applications are not greater than target response times of QoS classes corresponding to requests sent to the applications.

13. The method of claim 11, wherein the at least one server comprises a plurality of servers and each application runs on a different one of the plurality of servers.

14. Apparatus for processing a request to at least one server, comprising:

a memory; and

at least one processor coupled to the memory and operative to receive a request, and schedule submission of the request to the at least one server based on: (i) a quality-of-service (QoS) class assigned to a client from which the request originated; (ii) a response target associated with the QoS class; and (iii) an estimated response time associated with the at least one server.

15. The apparatus of claim 14, wherein the memory and the at least one processor form a scheduler that is external to the at least one server.

16. The apparatus of claim 15, wherein the scheduler is a front-end scheduler and the at least one server is a back-end server.

17. An article of manufacture for processing a request to at least one server, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

receiving a request; and

scheduling submission of the request to the at least one server based on: (i) a quality-of-service (QoS) class assigned to a client from which the request originated; (ii) a response target associated with the QoS class; and (iii) an estimated response time associated with the at least one server.

18. A method of processing requests to at least one server, comprising the steps of:

assigning at least one client to a quality-of-service (QoS) class from among at least two QoS classes;

assigning a response target to at least one QoS class;

estimating at least one response time of the at least one server based on one or more requests sent to the server within a given time period; and

withholding requests associated with a first one of the at least two QoS classes to allow requests associated with a second one of the at least two QoS classes to meet its response target based on the at least one estimated response time.

19. The method of claim 18, further comprising the steps of:

determining a throughput of the at least one server; and

reducing a request withhold rate to increase throughput of the at least one server.

20. The method of claim 18, further comprising the steps of:

monitoring a throughput of the at least one server; and

varying a request withhold rate to balance the throughput and request response times.

21. The method of claim 18, further comprising the steps of:

determining dispatch times for requests from a difference between at least one predicted response time of the at least one server and the target response time corresponding to the QoS class of the request; and

sending requests to the at least one server based on dispatch times.

22. The method of claim 18, wherein a plurality of applications are running on the at least one server and requests are routed to applications, further comprising the steps of:

estimating response times of applications based on one or more requests sent to the applications within a time period; and

sending a request to an application whose estimated response time is not greater than a target response time corresponding to the QoS class of the request.

23. The method of claim 22, further comprising the step of varying a number of requests sent to applications so that estimated response times of applications are not greater than target response times of QoS classes corresponding to requests sent to the applications.

24. The method of claim 22, wherein the at least one server comprises a plurality of servers and each application runs on a different one of the plurality of servers.

25. A method of providing a scheduling service for requests to at least one server, comprising the step of:

a service provider providing a scheduler operative to: (i) assign at least one client to a quality-of-service (QoS) class from among at least two QoS classes; (ii) assign a response target to at least one QoS class; (iii) estimate at least one response time of the at

least one server based on one or more requests sent to the server within a given time period; and (iv) withhold requests associated with a first one of the at least two QoS classes to allow requests associated with a second one of the at least two QoS classes to meet its response target based on the at least one estimated response time.